

*Review of:***A High-level Programming Environment for  
Packet Trace Anonymization and Transformation**

- **Paper by:**
  - Ruoming Pang: Dept. of CS / Princeton U.
  - Vern Paxson: International Computer Science Institute
- **Published in:**
  - SIGCOMM: Karlsruhe, Germany
  - August 2003
- **Review by:**
  - James Moscola

## Introduction

- **Packet traces of Internet traffic are great for research**
  - Most are not shared because they contain sensitive information
    - IP addresses, emails, personal web pages, etc.
  - Other traces do not contain packet payloads
    - Only contain TCP/IP headers with IP address renumbered
- **Paper describes tool that supports packet trace transformation**
  - Anonymize both packet headers and payloads
  - Can also be used for generic trace transformations
    - FTP, SMTP, HTTP, Finger, Ident

## Introduction

- Traces are processed as follows:
  - Payloads are reassembled and parsed to generate application protocol-level data elements
  - Data elements are processed by policy scripts to remove sensitive information
  - Data elements are converted back to bytes, frames, and packets

## Generic Trace Transformation

- Three steps:
  - Parsing
  - Data Transformation
    - Policy scripts
  - Composition

## Trace Parsing

- Flow reassembling
  - Reassemble IP fragments and TCP streams
- Line breaking
- Protocol-specific parsing
  - Currently has parsers for:
    - DNS, Finger, FTP, HTTP, ICMP, Ident, MIME, NTP, Netbios, Rlogin, SMTP, SSH, and Telnet

## Policy Scripts

- Written in the Bro language
  - Strong typing:
    - address, port, time, regular expression matching, string manipulation

## Policy Script Rewrite Example

- **Event**

```
smtp_request (  
  conn: connection = C,  
  command: string = "MAIL",  
  argument: string = "From: <alice@bob.org>")
```

- **Rewrite Rule**

```
rewrite_smtp_request (  
  C,  
  "MAIL",  
  "From: <name123@domain111>")
```

## Trace Composer

- **Rewrite Functions**
  - Currently have implemented: FPT, HTTP, SMTP, Finger, Ident
- **Packet Generation (Framing)**
  - How many packets
  - Timestamp
- **TCP/IP header fields**
  - Modifies the following fields:
    - IP addresses
    - Clears fragment bits
    - Give unique IP ID field
    - TCP sequence/ack fields are corrected
    - Checksums
    - Removes most options
    - Removes FIN flag (composer inserts its own)

## Trace Size Reduction

- Replaces HTTP entities beyond a specified size with MD5 hash value
  - 729 MB files with 0 bytes threshold
    - Reduces size to 25 MB
- Replaces email message bodies with MD5 hashes

## Trace Anonymization

- Information to hide
  - Identities
    - Users, hosts, etc.
  - Confidential attributes
    - Passwords, etc.
- Constant substitution:
  - Elements are no longer distinguishable
- Sequential Numbering:
  - File1, file2, file3 ...
  - Must maintain history ☹
- Hashing:
  - They are still unique

## Results

- Day-long trace (80 MB; 8,871 connections)
  - FTP analyzer : 131 seconds
  - FTP analyzer + anonymizer: 1009 seconds
  - FTP analyzer + dummy rewriter 192 seconds

## Work for the future

- Formalize security considerations
- Automate anonymization process
- Maintain traffic dynamics